Bollywood Cultural Trends Analysis

-By Kheya Ghosh Dastidar

The repository Bollywood_Trends_Analysis contains the code and data for a project collecting and scraping relevant data about Bollywood movies and doing a sentiment analysis. The goal is to classify movies by key social themes across several ideological dimensions.

Project Objectives

The key objective of this project is to identify key themes in Bollywood movies and classify them along corresponding ideological axes.

The following themes and axes have been utilised for the project –

Theme	Ideological Axis
Hindu–Muslim relations	Secular — Exclusionary
Gender Dynamics	Feminist — Misogynistic
Nationalism	Tolerant — Jingoistic
Caste Dynamics	Egalitarian — Casteist

Project Pipeline

1. Data Sourcing and Integration

- The primary dataset was sourced from The Indian Movie Database.
- Multiple files were merged to create a **master movie file** with comprehensive metadata and attributes.
- A sample of 100 movies released after 2010 was selected to ensure the thematic and sentiment analysis was current and consistent.

2. Collection of Movie Materials

a. Subtitle Files

- Initial efforts to collect ".srt" subtitle files via **OpenSubtitles** were blocked due to access restrictions.
- **Subscene** and **YIFY** were explored as alternatives, but English subtitle coverage was insufficient or scraping attempts were blocked.
- As a result, **subtitle data was not used** in the final analysis.

b. Plot Descriptions

- **Wikipedia** was the primary source for full plot descriptions.

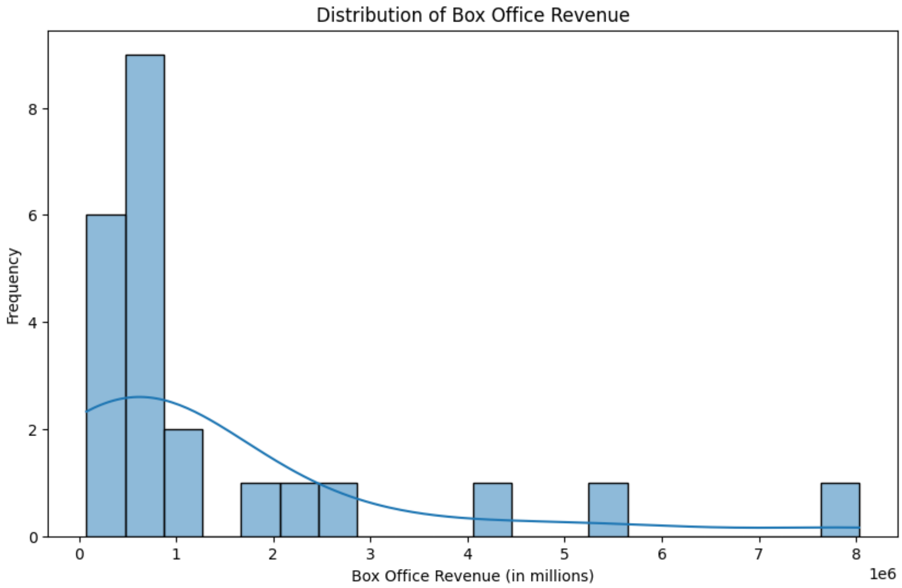
- If a Wikipedia plot was unavailable:
    - The project defaulted to the 'story' provided in the original dataset.
    - If neither was available, the **TMDb API** was used to fetch the official movie overview.
- This layered fallback system ensured the dataset remained robust while also maintaining efficiency and consistency.

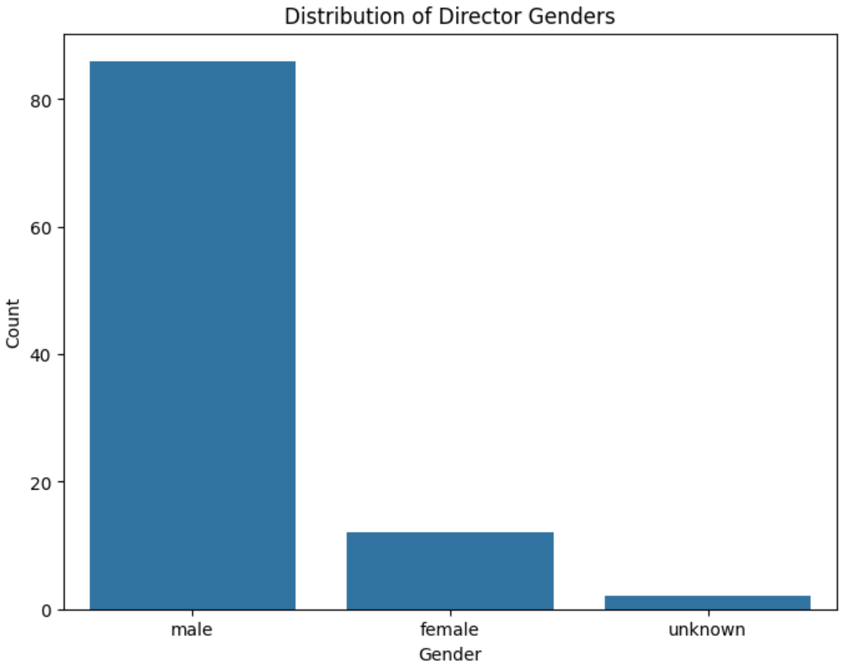
c. Visual Metadata

- Movie poster images were collected using the **TMDb API** for supplementary visual analysis.

d. Descriptive Metadata

- The **Directors' gender** was scraped from **Wikipedia** and inferred via the **Genderize.io API** where information was unavailable.
- **Box Office performance** was obtained using the **OMDB API**.
- Following are the summary plots for the same –





3. Thematic Classification

To measure the ideological sentiment embedded within Bollywood movie plots, a custom thematic classification pipeline was implemented using **zero-shot learning**.

Thematic Classification & Ideological Sentiment Mapping

Approach

- Each movie plot was evaluated on the following four major socio-political axes:
    - Hindu–Muslim relations: Secular - Exclusionary
    - Gender dynamics: Feminist - Misogynistic
    - Nationalism: Tolerant - Jingoistic
    - Caste representation: Egalitarian – Casteist
- A zero-shot classification pipeline from Hugging Face was used to assign the most likely label to each plot across these axes. This method allows classification without needing labelled training data.

Rationale and Model Choice

Initially, large language model APIs such as Groq and OpenAI were considered to better capture contextual nuance in longer plots. However, due to token size limitations and pricing constraints, they were not viable for a larger dataset (especially keeping in mind that the analysis should theoretically be able to be carried out on movie subtitle files).

Several Hugging Face models ("**facebook/bart-large-mnli**", "**roberta-large-mnli**") were also tested but caused session crashes and memory overflow errors due to their size and the amount of data.

Ultimately, "**cross-encoder/nli-distilroberta-base**" was selected for its lightweight architecture, compatibility with CPU and GPU environments, and reliable zero-shot performance without requiring fine-tuning.

Note: It is recommended to run the classification pipeline on ***Google Colab with GPU enabled*** for optimal performance. The code also runs on CPU, but with a significantly longer processing time.

Rationale for Chosen Axes

The chosen axes reflect key socio-political themes that have consistently appeared in Bollywood cinema and are central to public discourse in India. Each axis captures a dimension of ideological framing, making them well-suited for the analysis of a movie's positioning on contentious social issues.

Hindu–Muslim Relations (Secular - Exclusionary)

This axis reflects how films represent interreligious dynamics, especially in the context of India's increasingly polarized political landscape. Films may depict either inclusive, pluralistic relationships or align with majoritarian narratives. This axis captures communal framing and religious inclusivity.

Gender Dynamics (Feminist - Misogynistic)

This axis gauges whether films challenge or reinforce patriarchal structures. It is particularly relevant in evaluating female agency, objectification, and representation in mainstream cinema.

Nationalism (Tolerant - Jingoistic)

This axis evaluates whether patriotism is framed through inclusive, democratic ideals or through aggressive, jingoistic, and exclusionary forms of nationalism.

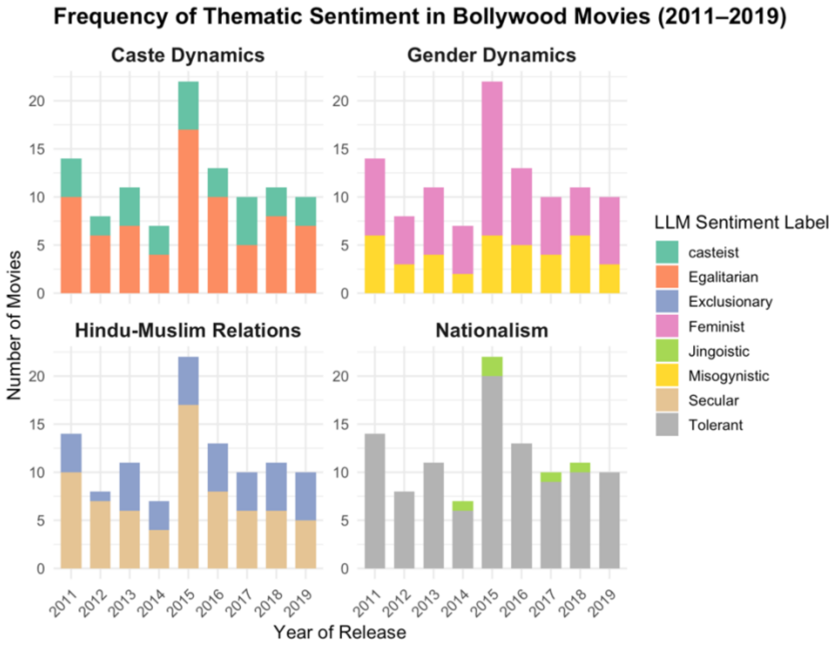
Caste Dynamics (Egalitarian - Casteist)

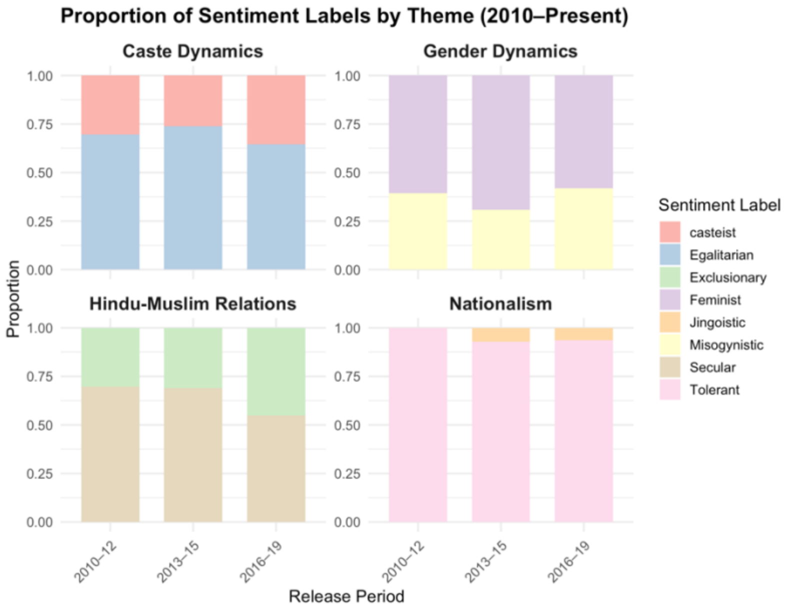
Given the centrality of caste in Indian society, it can serve as a critical lens to analyse cinema. This axis captures whether caste dynamics are portrayed in an equitable manner or whether stereotypical and regressive depictions are present or normalised.

Visualization & Analysis

- ggplot2 used for sentiment trend graphs over time
- matplotlib and seaborn used to create summary plots for the additional metadata

Following are the graphs showcasing how the frequency of each theme and its sentiment category changes over time and how the proportion of each sentiment category for each theme changes over the years.




Proportion of Sentiment Labels by Theme (2010–Present)
Caste Dynamics
Gender Dynamics
Hindu-Muslim Relations
Nationalism
Proportion
1.00
0.75
0.50
0.25
0.00
2010–12
2013–15
2016–19
Release Period
Sentiment Label
casteist
Egalitarian
Exclusionary
Feminist
Jingoistic
Misogynistic
Secular
Tolerant